Adaptive multi-class Bayesian sparse regression - An application to brain activity classification

Vincent Michel[1,2,5], Evelyn Eger[4,5], Christine Keribin[2,3], and Bertrand Thirion[1,5]

[1] Parietal team,INRIA Saclay-Île-de-France, Saclay, France ,
[2] Université Paris-Sud 11, Orsay, France,
[3] SELECT team, INRIA Saclay-Île-de-France, France
[4] INSERM U562, Gif/Yvette, France
[5] CEA, DSV, I2BM, NeuroSpin, Gif/Yvette, France

Abstract. In this article we describe a novel method for regularized regression and apply it to the prediction of a behavioural variable from brain activation images. In the context of neuroimaging, regression or classification techniques are often plagued with the curse of dimensionality, due to the extremely high number of voxels and the limited number of activation maps. A commonly-used solution is the regularization of the weights used in the parametric prediction function. It entails the difficult issue of introducing an adapted amount of regularization in the model; this question can be addressed in a Bayesian framework, but model specification needs a careful design to balance adaptiveness and sparsity. Thus, we introduce an adaptive multi-class regularization to deal with this cluster-based structure of the data. Based on a hierarchical model and estimated in a Variational Bayes framework, our algorithm is robust to overfit and more adaptive than other regularization methods. Results on simulated data and preliminary results on real data show the accuracy of the method in the context of brain activation images.

1 Introduction

Inferring behavioral information or cognitive states from activation brain images such as those obtained with functional magnetic resonance imaging (fMRI) is a recent approach in neuroimaging [**?**] that can provide more sensitive analyses than standard statistical parametric mapping procedures [**?**]. Specifically, it can be used to check the involvement of one or several brain regions in certain cognitive or perceptual functions by evaluating the accuracy of the prediction of a behavioral or cognitive variable of interest (the *target*) when the classifier is instantiated on that particular brain region. Such an approach is particularly well suited for the investigation of population coding [**?**]: certain neuronal populations are thought to activate specifically when a certain perceptual or cognitive parameter reaches a given value; inferring the parameter from the neuronal activity helps to *decode* the brain system.

The main difficulty of such a problem is the huge dimensionality of the data, with far more features (in our case, voxels) than samples (images). This problem

leads to overfit and thus dramatically decreases prediction accuracy. One common solution consists in working in the dual space using the kernel trick [], but in the case of neuroimaging, it may be more fruitful to compute explicit loadings on brain regions, hence to work in the primal space. To deal with this dimensionality problem, some regularized regression techniques have been developed, forcing the majority of the features to have zero or close to zero loadings, such as Lasso [**?**] and elastic net [**?**]; however, these approaches require that the amount of regularization is fixed beforehand, and possibly optimized by cross-validation. By contrast, Bayesian methods (e.g. adpative ridge Regression [] and Automatic Relevance Determination -ARD- [**?**]) adapt the amount of regularization to the problem at hand. These regularized regression methods have been already used for predicting cognitives states. In [**?**], a model based on ARD has been proposed for weighting activity patterns in the case of logistic regression, but ARD can overfit in the case of very high dimension. Similarly, in [**?**] a Bayesian regression approach is used to classify brain states, but the construction relies on ad hoc voxel selection steps, so that there is no proof that the solution is optimal. In summary, Bayesian regression techniques have been developed in two contexts: on the one hand, adpative ridge regression regularizes all the loadings with the same parameter, which is not adapted to brain activity where only few clusters of voxels have task-related activity; on the other hand ARD regularizes separately each voxel, and is prone to overfit when the model contains too many regressors.

In this article, we develop an intermediate approach for regularized sparse regression, which assigns each voxel to a class, the number of which is fixed by the operator. Regularization is perfomed in each class separately, leading to a stable and adaptive regularization, while avoiding overfit - this approach is thus a compromise between ridge regression and ARD. The algorithm is based on a Variational-Bayes (VB) approach which leads to a fast estimate of the weights distributions. The parameters update algorithm is no more complex than an Expectation maximization algorithm, and iteratively adapts the hyperparameters to our problem. Moreover, the VB approach has one important property for model selection : it contains a built-in criterion, i.e. the free energy of the model. After introducing our model and the VB approach, we show that the proposed algorithm performs better than reference methods for simulated data, and leads to promising preliminary results on real data.

2 Methods

We introduce the following regression model :

$$y = \Phi w + \epsilon \tag{1}$$

where y represents the behavioural data to be fit (written as a vector of length n) and w the parameters (written as a vector of length m). Φ is the design matrix and can be a $n \times m$ matrix (each row is an m-dimensional activation map), or possibly a $n \times n$ matrix (in the case of use of a bilinear kernel). The main problem with this model is that $n \ll m$, so that estimating w is an ill-posed

problem. A solution is to introduce some priors over the parameter distribution.

Priors on regression and adaptative relevance determination (ARD)
Regularized regression can be used to solve ill-posed problem, by imposing a prior on the weights, hence possibly a sparse feature weighting. First, we model the noise with a Gaussian density:

$$\epsilon \sim \mathcal{N}(0, \sigma^2 I_n) \tag{2}$$
$$p(\sigma^2) = \Gamma^{(-1)}(\lambda_1, \lambda_2) \tag{3}$$

with an hyperparameter σ. $\Gamma^{(-1)}$ stands the inverse gamma distribution.
The prior on the weights is given by a gaussian prior :

$$w \sim \mathcal{N}(0, A^{-1}) \tag{4}$$
$$p(\alpha) = \Gamma(\gamma_1, \gamma_2) \tag{5}$$

where $A = \mathrm{diag}(\alpha_1, ..., \alpha_m)$, the hyperparameters $\alpha_i, i \in [1, m]$ are the precision parameters, and Γ is the gamma density. Two important cases correspond to adaptive ridge regression ($\alpha_1 = ... = \alpha_m$) and ARD ($\alpha_i \neq \alpha_j$ if $i \neq j$). Still, the highly adaptive regularization of the ARD can lead to severe overfitting if $n \ll m$.

Multi-Classes model (VBK)
In order to accommodate the sparsity of ARD with the stability of adaptive ridge regression, we introduce an intermediate representation, in which voxels belong to one class among K following the discrete variable z. Thus all the features within a class $k \in [1, .., K]$ share the same precision parameter α_k. This complete generative model is summarized in Fig.1. Next, we introduce a prior for z :

$$p(z_i) = \prod_{k=1}^{K} \pi_k^{z_i = k}$$

and a Dirichlet prior on π_k with hyperparameter δ: $p(\pi) = Dir(\delta)$

Estimation and Selection of the model by Variational Bayes
To select a model among several alternatives, it is natural to keep the model that yields the largest data evidence $p(y)$, which needs to be approximated. Thus, we use the variational approximation which allows us to find a closed-form $q(\theta)$ of $p(\theta|y)$, where $q(\theta)$ is in a given family of distributions and $\theta = [\sigma^2, z, \alpha, w, \delta]$ are the parameters of the model. We can then decompose $\log p(y)$ as the sum of the free energy $\mathcal{F}$ and the Kullback-Leibler divergence between the true posterior $p(\theta|y)$ and the variational approximation $q(\theta)$:

$$\log p(y) = \mathcal{F}(q(\theta)) + D_{KL}(q(\theta)||p(\theta|y)) \tag{6}$$
$$\mathcal{F} = \int \log \frac{p(\theta, y)}{q(\theta)} q(\theta) d\theta \tag{7}$$

Thus, the free energy $\mathcal{F}$ is a lower bound of $\log p(y)$ with equality iff $q(\theta) = p(\theta|y)$, and inferring the density q of the parameters corresponds to maximizing $\mathcal{F}$. Moreover, free energy is a measure of the quality of the model and can be used in a model selection scheme and avoids the global time-consuming cross-validation-based optimization of K.

Initialization and validation
The initialization is set as [**?**], with weakly informative prior, $\lambda_1 = \lambda_2 = \gamma_1 = \gamma_2 = 10^{-6}$ and $\delta = 5$ (see [**?**]). Since the estimation algorithm converges to a local maximum of $\mathcal{F}$, the algorithm is very sensitive to initialization of z, performed by using a K-Means on the F-statistics of the features. Then we developed a two-steps procedure to avoid this problem : (i) with z fixed, update all the parameters except z, $q(z)$ and δ until the convergence of all the α_k. (ii) z is no longer fixed, and all the parameters are updated until the convergence of all the α_k.

The performance of the competeing models is computed using the ratio of explained variance ζ. Let Φ^l, Y^l be a learning set, Φ^t, Y^t a test set, and $\hat{Y}^t(\Phi^l, Y^l, \Phi^t)$ the prediction obtained with a model trained Φ^l, Y^l and tested with Φ^t.

$$\zeta(\Phi^l, Y^l, \Phi^t, Y^t) = \frac{\mathrm{var}(Y^t) - \mathrm{var}\left(Y^t - \hat{Y}^t(\Phi^l, Y^l, \Phi^t)\right)}{\mathrm{var}(Y^t)} \tag{8}$$

This is the amount of variability in the response that can be explained by the model (prediction is perfect if $\zeta = 1$, and is worst than chance if $\zeta < 0$).

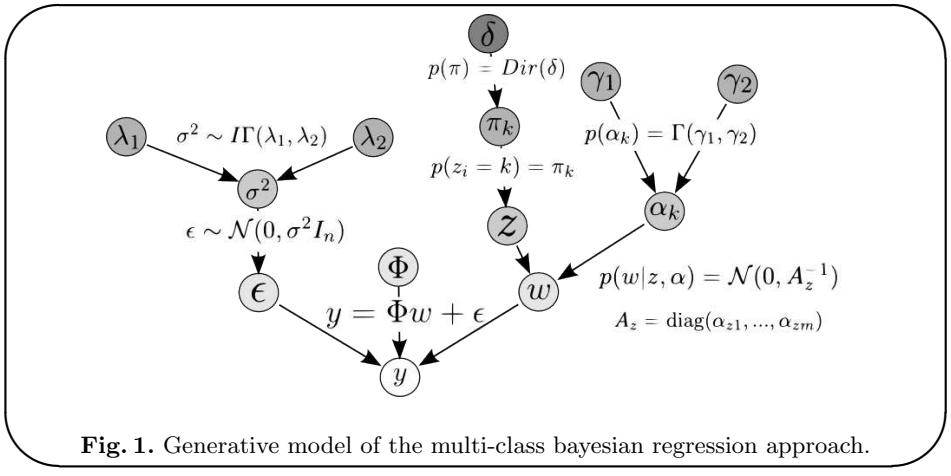


Fig. 1. Generative model of the multi-class bayesian regression approach.

3 Results

3.1 Simulated Data

We tested our algorithm on a simulated data set X of N_p images with squared Regions of Interest (ROIs) $\mathcal{R}$ (defined by a position and a width). We note b the background (i.e. outside the ROIs). The signal in the (i, j) voxel of the k^{th} image is simulated as :

$$X_{i,j,k} = \sum_{r \in \mathcal{R}} \mathbb{I}_r(i,j)\alpha_{r,k}u_{i,j,k} + \mathbb{I}_b(i,j)u_{i,j,k} + \epsilon_{i,j,k} \tag{9}$$

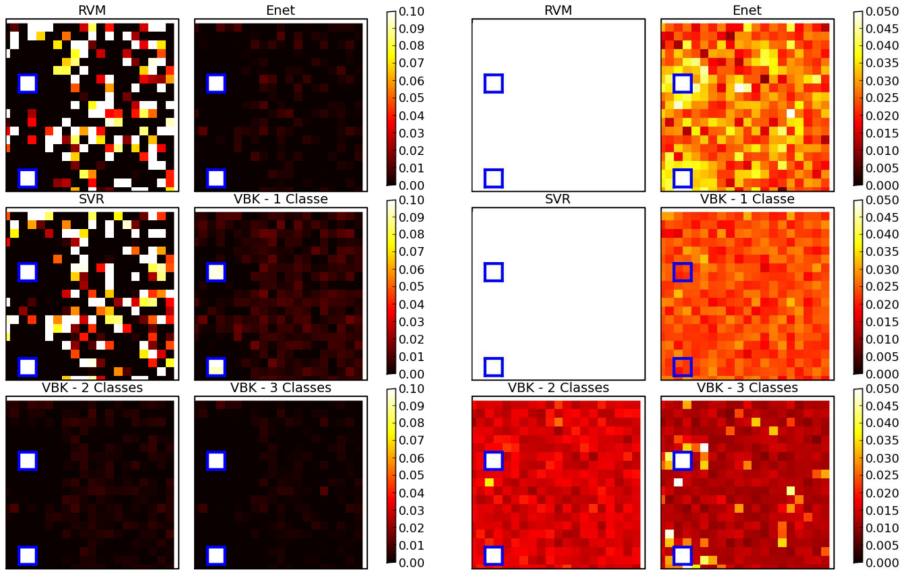
where $u_{i,j,k}$ is a random value from an uniforme distribution in $[0, 1]$, $\epsilon_{i,j,k}$ a random value from a Gaussian distribution $\mathcal{N}(0, 1)$ smoothed with a parameter of 2 voxels to mimic the correlation structure observed in real fMRI datasets, $\alpha_{r,k} \sim \mathcal{U}[0, 1]$ for ROI r and image k. We have $\mathbb{I}_r(i, j) = 1$ (resp. $\mathbb{I}_b$) if the (i, j) voxel is in r (resp. b), and $\mathbb{I}_r(i, j) = 0$ (resp. $\mathbb{I}_b$) elsewhere. We simulate the target Y as : $Y_k = \sum_{r \in \mathcal{R}} \alpha_{r,k}$
We generate a a dataset of 250 images, and split it in a learning set of 200 images and validation set of 50 images. The images have a size of 20×20, with two non-overlapping ROIs of width 2 pixels. We compare our algorithm with three other methods : a bilinear kernel-based ARD regression (also called Relevance Vector Machine *RVM* [**?**]), an elastic net regularization procedure (which we will call *Enet* [**?**] with parameters $s = 0.5$ and $\lambda = 0.1$), and a Support Vector Regression procedure (which we will call *SVR* [**?**] with a linear kernel and $C = 1$).

3.2 Results on Simulated Data

In a first experiment, we average the results obtained for the different methods for 40 tests. We test the values 1, 2 and 3 for the parameter K. See the results Fig.2: the *VBK* algorithm outperforms the other methods for $K > 1$ (c). Moreover, the *VBK* method finds very low and stable weights outside the ROIs (a,b), where *Enet* leads to a sparse (many weights are closed to zero) but less stable (higher standard deviation) regularized solutions. Both *RVM* and *SVR* yield a poorly regularized solution : many irrelevant voxels have a significant weight.
In a second experiment, we compute the explained variance and the free energy for different models with $K \in [1, 2, 3, 4, 5]$ for 20 samples (see Fig. 3). The model with the lowest free energy ($K = 1$) is the one with the worst prediction accuracy. We can see that the increase of free energy is strongly correlated with the increase of explained variance, with a maximum reached for $K = 3$.

3.3 Real data

We use real dataset related to a *numerotopy* (mental representations of quantities) experiment. During the experiment, ten healthy volunteers (6 males and 4 females, mean age 21.2 +/- 3.0 years) view dot patterns with different quantities



(a) (b)

(c)

	VBK - K=1	*VBK - K=2*	*VBK - K=3*	*RVM*	*Enet*	*SVR*
average ζ	0.05	0.35	0.39	0.09	0.31	0.16
std. deviation ζ	0.16	0.18	0.13	0.17	0.18	0.17

Fig. 2. Results of the simulation experiment. ROIs are outlined by blue squares. Mean (a) and standard deviation (b) for the weights found with different methods. The *VBK* approach gives weights similar to those of the *Enet* method, but with more stable estimation outside the ROIs. The *RVM* and *SVM* approaches lead to non-zero weights outside the ROIs, and weights estimation is not stable across trials. (c) Ratio and standard deviation of the explained variance for different methods averaged on 40 simulations. The *VBK* algorithm outperforms all the other techniques and yields less variable results (when $K > 1$).

of dots ($\nu = 2$, 4, 6 and 8; we take $Y = \log(\nu)$), with 8 repetitions of each stimulus : so that we have a total of $N_p = 32$ images by subjects. Functional images were acquired on a 3 Tesla MR system with 12-channel head coil (Siemens Trio TIM) as T2* weighted echo-planar image (EPI) volumes using a high-resolution EPI-sequence. 26 oblique-transverse slices covering parietal and superior parts of frontal lobes were obtained in interleaved acquisition order with a TR of 2.5 s (FOV 192 mm, fat suppression, TE 30 ms, flip angle 78° , $1.5 \times 1.5 \times 1.5$ mm voxels). Standard pre-processings and the fit of the general linear model have been performed with the SPM5 software. Signals magnitude are expressed as percentage of increase compared to the baseline.

We keep 1000 voxels included in the main region of interest, i.e. the Intra-Parietal Sulcus (IPS), which has been manually delineated in all the available datasets

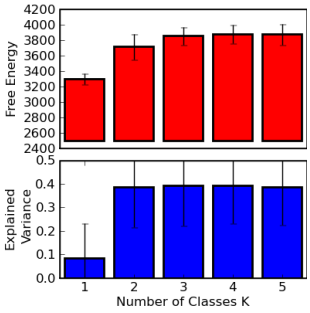


Fig. 3. Results of the simulation experiment for model selection. The free energy (red) and the explained variance (blue) are average for 20 simulations, and are strongly correlated, with a maximum reached for $K = 3$. Thus, the free energy of the *VBK* model can be used for the selection of the model.

prior to fMRI data analysis. Thus, in order to further reduce the dimensionality of the data, we parcellate this region in 200 parcels with a variant of Ward's algorithm, and we average the signal within each parcels.

3.4 Results on Real data

We compute the explained variance obtained in a leave-one-session-out procedure for different values of K with the *VBK* algorithm. We average the results across the 10 subjects. See Fig.**??**: (a) Example of loadings found by the *VBK* algorithm for one subject superimposed on the anatomical image. We can see that the *VBK* provides explicit weighting maps that allow to understand the anatomical orgnaization of discriminant brain activity. (b) free energy (top) and explained variance (bottom) averaged across subjects for different values of K. They are strongly correlated, and increasing the number K of classes in the model decreases the explained variance. This means that the proposed approach favors sparse parameterizations.

4 Discussion

Regularization of voxels loadings significantly increases the prediction accuracy. However, this regularization has to be adapted to the specific nature of each particular fMRI dataset, which is done in this article by introducing a multi-class framework. Our approach performs an adaptive and efficient regularization, and is a compromise between a global regularization (ridge regression and Lasso) which does not take into account the region-based structure of the information, and a the (ARD) which is subject to overfit when used in the primal space.

On simulated data, our approach performs better than other classical methods such as *SVR*, *RVM* and *Enet*. Besides an increase of the explained variance which shows that the *VBK* approach extracts more information from the data, the loadings are less noisy and more stable, leading to more interpretable activation maps. The correlation between the free energy and the prediction accuracy confirms that free energy is a valuable model selection tool that furthermore avoids time-consuming optimization by cross-validation.

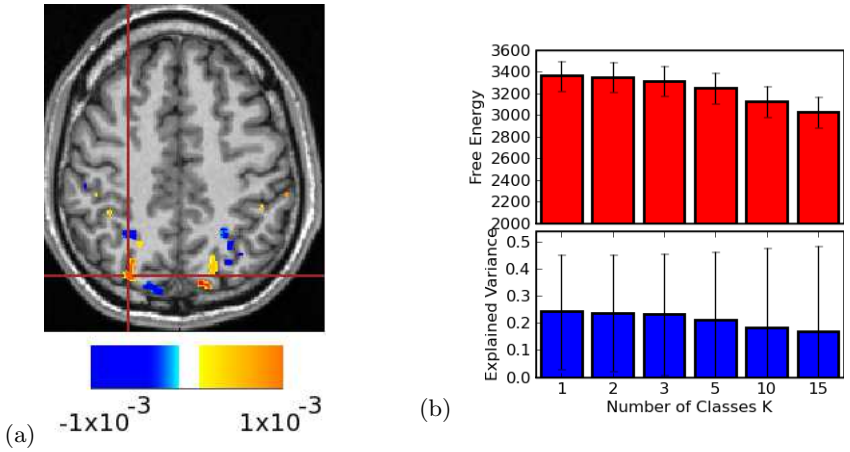


Fig. 4. Results on the real data for model selection. (a) Weights of the parcels found by *VBK* algorithm and superimposed on the anatomical image. Our method gives promising results for interpretative activity maps. (b) The free energy (red) and the explained variance (blue) are average for the 10 subjects, and are strongly correlated. Thus, in these preliminary results on real data, the free energy of the *VBK* model can be still used for the selection of the model.

Preliminary results on real data shows the advantages of our method. The *VBK* algorithm gives access to highly interpretable loadings maps which are a powerful tool for understanding brain activity. Moreover, the free energy seems to be an accurate built-in criterion for model selection. A future direction of our work is to optimize the spatial model used in our framework (here we simply use a prior parcellation of the search volume) in relationship with the prediction function taht we use. In parellel, we will develop non-linear versions (e.g. logistic/probit) of this model for classification.

Conclusion We have presented a multi-class regularization approach that includes adaptative ridge and automatic relevance determination as limit case; the ensuing problem of optimizing the number of classes is easily dealt with in teh Variational Bayes framework. Our simulations and experiments on real fMRI data show that this approach is well suited ofr neuroimaging, as it yields a powerful framework but also reliable and interpretable feature loadings.

References

1. Cox, D.D., Savoy, R.L.: Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage **19**(2) (2003) 261–270
2. Kamitani, Y., Tong, F.: Decoding the visual and subjective contents of the human brain. Nature Neuroscience **8**(5) (April 2005) 679–685

3. Dayan, P., Abbott, L.F.: Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. The MIT Press (2001)
4. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B **58** (1996) 267–288
5. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society B **67** (2005) 301–320
6. Bishop, C.M., Tipping, M.E.: Variational relevance vector machines. In: UAI '00: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence. (2000) 46–53
7. Yamashita, O., aki Sato, M., Yoshioka, T., Tong, F., Kamitani, Y.: Sparse estimation automatically selects voxels relevant for the decoding of fmri activity patterns. NeuroImage **42** (2008)
8. Friston, K., Chu, C., Mourao-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J.: Bayesian decoding of brain images. NeuroImage **39** (January 2008) 181–205
9. Tipping, M.: The relevance vector machine. In: Advances in Neural Information Processing Systems, San Mateo, CA. (2000)
10. Penny, W., Roberts, S.: Variational bayes for 1-dimensional mixture models. (2000)
11. Cortes, C., Vapnik, V.: Support vector networks. In: Machine Learning. Volume 20. (1995) 273–297